January 2021

Perspective
EXPERT INSIGHTS ON A TIMELY POLICY ISSUE

DOUGLAS YEUNG, INEZ KHAN, NIDHI KALRA, OSONDE A. OSOBA

Identifying Systemic Bias in the Acquisition of Machine Learning Decision Aids for Law Enforcement Applications

Biased software tools that use artificial intelligence (AI) and machine learning (ML) algorithms to aid in decisionmaking ("decision aids") can exacerbate societal inequities when used in some domains, such as health care (Obermeyer et al., 2019) and criminal justice (Angwin et al., 2016). During a time of civil unrest, it is crucial to ensure that such decision aids—in particular, those used by law enforcement (LE) agencies—produce equitable outcomes. For example, the U.S. Department of Homeland Security (DHS) already fields such decision aids (e.g., facial recognition for airport screening [Oliver, 2019]) and is considering others (DHS, 2018). These current and planned software implementations should be examined for potential bias.

Existing efforts to address AI bias typically focus on how ML models are developed and trained. But bias also can creep in at other steps in software

RAND CORPORATION

implementations, beginning with acquisition and continuing to deployment. This Perspective describes an initial effort to examine other potential entry points for bias in ML decision-support tools and to identify opportunity areas to improve the use of ML tools in LE applications.

Algorithmic decision aids (i.e., tools that rely on AI and ML models) are increasingly a part of LE operations. These technologies have the potential to be deployed widely, such as in biometric systems (e.g., facial recognition) or for predictive policing (e.g., mapping crime hot spots, predicting risk of a person becoming involved in a violent or serious crime). As a result, it is important to understand any challenges that exist in equity, efficiency, and effectiveness of these individual systems. Furthermore, the manner in which these systems are integrated into decisions can systematically amplify adverse impacts of any of these challenges. For example, algorithmic outcomes that are uncritically accepted as impartial could lend unjustified authority to inequitable decisions.

Biased use of AI and ML applications could undermine what previous RAND Corporation work has described as *legitimacy policing* (Hollywood et al., 2018). The issue of legitimacy in LE is particularly important given the renewed focus on racial injustice in the wake of protests after George Floyd was killed during an arrest in May 2020. Although many police officers work extremely hard to maintain people's safety, public trust in police has fallen to record low levels (Brenan, 2020), while racial disparities in LE (e.g., arrests, use of force) persist (Williams, 2016). A key tenet of legitimacy policing is that procedural justice is crucial to the legitimacy of LE. Procedural justice seeks to ensure that civilians perceive that their interactions with police are fair and that their voices are heard, regardless of outcome. Another major part of legitimacy policing is dialogue with the community. Citizen participation could be a critical way to address bias in AI and ML, which, in applications used by LE, could undermine "core ingredients of procedural justice" (Mazerolle et al., 2013), such as perceptions of neutrality, the treatment of people with dignity and respect, and the trustworthiness of LE's motives. Policing experts view good community relations not only as a way to improve legitimacy but also as a core objective of policing itself. This is because people are more likely to follow the law and cooperate with the police when they perceive the law and LE as having legitimate authority (see Hollywood et al., 2018, for details).

Abbreviations

AI	artificial intelligence
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
DHS	U.S. Department of Homeland Security
FEMA	Federal Emergency Management Agency
LE	law enforcement
ML	machine learning
TSA	Transportation Security Administration

How Machine Learning Supports Decisionmaking

ML models are statistical models that can be deployed for various decisionmaking contexts, such as in DHS operations, to provide consistency and scalability. Most such models fall into one of three broad categories: supervised, unsupervised, and reinforcement learning models. Of these, supervised ML models are the most directly applicable to operational decisionmaking contexts, such as facial recognition or biometrics. In deployment, these ML models might be best conceptualized as question-answering artifacts. Training for these models consists of learning consistent and accurate behavior by examining past data or past examples of question-and-answer pairs. Once trained, a model should perform well during deployment if both of the following conditions hold:

- It has **contextual regularity:** The decisionmaking context or environment remains identical between training and deployment; (i.e., "don't use a model built for one decision for a separate decision, however related").
- It has **statistical regularity:** The population samples on which decisions are to be made are statistically identical during both the training and deployment phases (i.e., "don't use a model built for one population on a different population").

Accurate and consistent ML models help to improve or streamline mission-critical operations. But all decisionmaking processes, whether made by humans or by ML models, contain some inherent flaws, which might be most salient and keenly felt when decisionmaking outcomes directly affect humans. Bias is a particularly important flaw in decisionmaking processes based on statistical ML models.

What Is Bias?

Given the increasing awareness of bias in many facets of society, the question of what constitutes bias might seem self-evident. Yet a serious look at this question reveals that defining bias is a fraught endeavor (Osoba, Boudreaux, Saunders, et al., 2019). There is no single definition of either *bias* or *equity*. This is important because equity attempts to account for systemic or contextual factors, such as bias. Three problems arise with any attempt to conceptualize equity. First, equity is not a singular concept. Different equity norms often apply in different situations. Second, seemingly reasonable concepts of equity can be contested or even incompatible with each other. Third, prescriptive and theoretical

Once trained, a model is theoretically guaranteed to perform well during deployment if it is both contextually and statistically regular.

Governments may be held to fairness standards that differ from those applied to the private sector.

concepts of equity sometimes differ from common practical concepts.

The debate around the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) criminal risk assessment tool offers a compelling illustration of this problem.[1] COMPAS uses a large set of factors to assess the risk that a person will commit another crime following release from custody. In 2016, ProPublica published an analysis that indicated that COMPAS was **twice as likely** to mischaracterize black people as high risk than it was white people (Angwin et al., 2016). This highly unequal error rate and its corresponding effect on black people grounded a compelling argument that the system was racially biased. The makers of COMPAS argued, however, that it was not racially biased (Dieterich, Mendoza, and Brennan, 2016). Alexandra Chouldechova mathematically proved that an accurate classification algorithm will produce error rates based on the makeup of the data being classified (e.g., if two populations are unequally represented in a data set, the algorithm will produce unequal error rates) (Chouldechova, 2017). Thus, the makers argued that COMPAS was not biased and that the unequal error rates were a natural consequence of a world in which black people are disproportionately incarcerated. Does the fact that COMPAS was twice as likely to mischaracterize black people than white people as high risk mean that the model is biased? Or are the data that feed the model—coming from parts of the underlying criminal justice system—biased? And will the use of COMPAS reduce or entrench those biases? Is bias of greater concern for punitive uses of algorithms? These questions demonstrate how differently people can view the presence of bias within a single algorithm and its differential impact.

Further, the characteristics by which we identify the groups for disparate-impact analyses are typically subjective choices based on social choices, legislative processes, or even historical accident. For instance, Title VII of the Civil Rights Act of 1964 (Pub. L. 88-352) protects employees from discrimination based on specified characteristics: race, color, national origin, sex, and religion. This legislative mandate justifies disparate-impact analyses on the basis of these demographic characteristics. Yet even this legislative mandate was not immutable—a recent landmark decision from the U.S. Supreme Court interpreted "sex" to include gender identity and sexual orientation, a broadening of the earlier interpretation of this statute (Liptak, 2020).

Governments may also be held to fairness standards that differ from those applied to the private sector, reflecting the concern that bias causes disproportionate harm to vulnerable people. For instance, it is routine for a private company to make products affordable to only the wealthiest individuals, while governments might need to build products (e.g., infrastructure) that are

accessible to all. Although the government might face challenges in having to build accessible products, vulnerable groups are likely to benefit in the long run from having their needs broadly considered.

For this exploratory work, we settled on a simple, nonexhaustive definition of *bias*. We deem an automated decisionmaking system to be biased if (1) it consistently produces disproportional outcomes for different groups of people and (2) the disparate impacts are not commensurate with what might be expected for people in the affected groups given their relative proportion of the population.

Why Is Bias in Machine Learning Important?

The goal of this effort is to help DHS understand why careful consideration of bias in operational ML is important and how to frame its thinking as it tries to deploy more ML-based decision products in various missions and operations. Our aim is *not* to define what bias in operational ML means, given the broad variety in the operations of DHS and its components.

Given the normative complexity of trying to understand bias in decisionmaking, why should DHS (or any other mission-driven government agency) care about the question of ML-based bias in its operational decisionmaking contexts? There are three compelling and interconnected reasons to take bias analyses seriously:

- **a legislative mandate:** Some operations, missions, and institutions are legally mandated to achieve specific forms of equity. This is the simplest and clearest motivation for examining bias in ML applications. In the best case, there is a clear statement of what *equity* means in that context, how to measure it, and mechanisms for maintaining accountability for those equity constraints. Examples of such legislative mandates include the due process and equal protection clauses of the U.S. Constitution, which apply to U.S.-based

An automated decisionmaking system is biased if (1) it consistently produces disparate or disproportional outcomes for different groups of people and (2) the disparate impacts are not commensurate with what might be expected for people in the affected groups.

criminal justice institutions, and Equal Employment Opportunity Commission– and Title VII–style mandates that apply to U.S.-based employment institutions. The computer matching provisions of the U.S. Privacy Act offer an example of legislation that could apply to ML technologies, such as biometrics, while predictive policing technologies may require additional legislation (Ferguson, 2017).

- **an ethical mandate:** Operational decisions can be judged by widely shared equity norms based in ethics. This influence is a form of social or cultural coercion and might seem like a comparatively weak mandate, but it is an important and relevant form of coercion for government institutions in free civil societies because democratic institutions are, by definition, intended to be expressions of popular will. Government institutions that consistently violate ethical mandates in a free society might receive extensive pushback from the populace.
- **an operational mandate:** Efficient, equitable decisionmaking wastes less of the scarce resources of time, attention, and materiel that would be spent in mitigating or reversing biased decision processes with poor results. In a surveillance operation, for example, a biased decision process might contribute to spending more time than necessary on bad leads, thereby having fewer resources to spend on leads that actually pose legitimate security threats. Further, it can run the risk of sowing mistrust in the unfairly targeted subpopulation in the long term.

Although addressing bias in individual ML models is important, focusing on the flaws of individual models is insufficient to guarantee that operational outcomes are free from bias. Recent work (e.g., Osoba, Boudreaux, and Yeung, 2020; Raji et al., 2020) highlights the need for a more system-level perspective to address the presence of bias in overall decisionmaking missions and institutions.

Evaluating ML models on the basis of bias or equity represents a dimension of assessment that can diverge from the goal of making models as accurate as possible overall. Training models to be both accurate overall (for an entire population) and equitable (similarly accurate for relevant subgroups) may require constrained optimizations where those objectives partially conflict with one another (Donini et al., 2018). This potential divergence between equity and overall accuracy suggests that ML model development practices should be updated to explicitly quantify trade-offs where they occur. Aiming to optimize the separate objectives of equity and overall accuracy can lead to tension, depending on the decisionmaking context. One difficult scenario occurs when the decision outcome in question is strongly correlated with identifiers of statutorily protected categories (U.S. National Archives and Records Administration, 2016).

What We Did

We designed our method to represent the initial step in a process to help LE agencies identify up-front potential for bias and disparate outcomes *before* beginning the software acquisition process. Thus, we sought to identify key points in the LE software acquisition life cycle at

which bias might occur and to develop ways to evaluate which actions could help mitigate the impact of that potential bias. Drawing on RAND research and discussions with RAND experts, we first developed a notional acquisition framework. This framework consists of steps that are intended to collectively represent the overall life cycle of a software acquisition process.

Using this acquisition framework and drawing on previous RAND work on algorithmic audits, we examined each acquisition step for different types of bias and considered the set of possible actions to detect and mitigate them. We reviewed options for addressing AI and ML bias in LE systems and spoke with DHS subject-matter experts to assess how such a framework could be most useful.

Using the synthesized results, we identified opportunity areas ripe for further study and developed approaches that DHS could consider using to help identify how to safeguard against bias in automated decisions or mitigate the adverse systemic impacts of algorithmic decision aids. This Perspective describes these approaches, along with the acquisition framework, with supporting text that illustrates key points that DHS should audit for bias in software deployment. Finally, we discuss opportunity areas for future actions or research that DHS could undertake.

A difficult scenario occurs when the decision outcome in question is strongly correlated with identifiers of statutorily protected categories.

Identifying Sources of Bias in the Acquisition of Machine Learning Tools

To begin thinking about how bias could affect DHS's acquisition of ML tools, we lay out a notional acquisition framework consisting of five steps, as shown in the figure:

1. acquisition planning
2. solicitation and selection
3. development
4. delivery
5. deployment, maintenance, and sustainment.

This list of steps does not depict the full scope of DHS acquisitions but illustrates the main steps in the process at which ML bias concerns might emerge. This framework was created after reviewing existing DHS and U.S. Department of Defense documentation and regulations on acquisitions, including the Federal Acquisition

A Notional Acquisition Framework

Regulation. Future work will need to develop a more detailed and more accurate version of this framework.

In this Perspective, we argue that, of the five steps in the acquisition process, the steps in which the acquired system is created (development) and used (deployment, maintenance, and sustainment) are the principal ones in which bias is introduced. However, the other three steps—acquisition planning, solicitation and selection, and delivery—can influence whether and the extent to which development and deployment, maintenance, and sustainment are biased. Therefore, addressing bias needs to be a life-cycle concern that is baked into the acquisition process from the very beginning and not something that can be addressed effectively as an afterthought.

In this section, we describe each step of our notional acquisition framework, and we suggest several pathways that could result in bias in a final deployed AI or ML product. This description is not meant to be conclusive or exhaustive but to serve as a starting point for further understanding of processes by which bias can be introduced into a system and how bias can be identified and mitigated. (See the appendix for additional examples of types of bias that can arise in each stage.)

Acquisition Planning

As defined in the Federal Acquisition Regulation,

> *Acquisition planning* means the process by which the efforts of all personnel responsible for an acquisition are coordinated and integrated through a comprehensive plan for fulfilling the agency need in a timely manner and at a reasonable cost. It includes developing the overall strategy for managing the acquisition. (48 C.F.R. § 2.101)

The first stage in the acquisition process is identifying a need for a product or service. From here, a plan is developed to detail the item's proposed use and requirements. Details of the plan include explaining the problem, identifying how the proposed product or service provides a solution, laying out the high-level design and functionality of the proposed product, and defining specific requirements and caveats for the proposed product or service. Because the acquisition planning process frames the rest of the acquisition life cycle, we hypothesize that the doors to bias can be opened at this stage in several ways and that, therefore, bias mitigation efforts must begin here.

First, we suggest that it is important for the acquisition team to fully understand the social context in which the acquired technology will be used and ways

in which that social context might itself be biased. This context can lead to bias in the system's use, even in the absence of bias in the ML model itself. For example, consider COMPAS. Studies show that incarceration rates, for a variety of reasons, are disproportionate by race (National Research Council, 2014). For example, more black people are incarcerated relative to how many black people are in the population. When these data are fed into the COMPAS tool, the COMPAS algorithm makes disproportionate recidivism predictions by race. That is, when the underlying rates of incarceration for black and white people are disproportionate to their demographic distributions, the risk tool makes disproportionate risk classification error rates by race, even if the tool might not itself have made disproportionate errors.

Second, the use case for which the technology is intended and its user interface might be mischaracterized in terms of whom they will affect and how they might differently affect people with different characteristics. A mischaracterization can lead to violations of contextual regularity, yielding a technology that is not developed for the circumstances in which it is actually used, with a potential for biased results in the actual target population.

Third, concerns about bias are at least partly motivated by the fact that judgments from an ML system may often be used to assign punitive consequences, for which errors can cause great harm, such as through misidentification. Concerns about bias are motivated in part by a recognition that such harm ought not be disproportionately borne by certain groups of people, particularly those who have historically experienced discrimination or who are already vulnerable.

Errors in an ML system can cause great harm, and such harm ought not be disproportionately borne by groups of people who have historically experienced discrimination or who are already vulnerable.

The harm from errors should be deliberately identified in specific ML applications. Imagine that an LE agency uses a facial recognition system in an attempt to identify and apprehend a suspected domestic terrorist, a situation that could involve deadly force. An error in the facial recognition system can lead to the ultimate cost: unjust loss of life. There might be more consensus around and more urgency in fighting bias in such a system.

However, in other cases, there might be less consensus about what constitutes harm. In an immigration-overstay risk tool, for example, a determination that someone is high risk can result in that person being monitored to ensure that they do not overstay

their visa. To some, such monitoring could seem like a small consequence, not meaningfully different from *not* being monitored. Under such a view, the need to address bias might be less urgent because the harm from an incorrect determination about a person's risk might be perceived as insignificant. In other words, the cost of the tool being wrong might seem small. But ongoing monitoring of entire populations deemed to be high risk for overstays could be seen as unacceptable to those populations and to civil rights and civil liberties groups. This implicit value judgment could result in less vigilance and more bias than if monitoring were seen as invasive and in violation of an individual's rights (i.e., under a view in which errors are very costly). The acquisition planning process should include deliberation over the human impact of different outcomes, make explicit all subjective judgments about those differences, and embrace a multiplicity of values and opinions. The results of these deliberations should inform requirements and provisions on mitigating bias in capability needs, concepts of operations, requirements, and design documents, such as in the Joint Requirements Integration and Management System process.

Finally, research has shown that bias may be reduced when the team making the decisions is representative of the affected populations, in that the team members would presumably have a better understanding of those populations because of the characteristics team members share with them (Todd et al., 2011). Conversely, a team whose members share few characteristics or lack familiarity with those who will be affected—particularly those who might be negatively affected—might not recognize bias or its harms. The composition and characteristics of members of the acquisition planning team could shape what biases can get introduced or mitigated. For example, diverse work teams, particularly those with "culturally competent" leaders, more effectively share and analyze organizational information (Groves and Feyerherm, 2011), which could aid in recognizing bias. Thus, having an acquisition team that is representative of the population targets for whom the technology is intended should lead to better outcomes.

Solicitation and Selection

A key follow-on step is to solicit proposals for development of the technology from vendors and to select one or more of those vendors as providers. Proposals can be acquired by issuing requests to vendors, by receiving bids from them, or both. After comparing the bids that have been received, the component can then award a contract based on a variety of criteria, including responsiveness, capability, cost, and timeliness.

The solicitation and selection of vendors to deliver an AI or ML application can also open the door to bias. First, the language in the solicitation, or the criteria by which vendors are selected, could be biased or favor certain respondents over others, even potentially discouraging some vendors from applying and resulting in a less-diverse vendor pool that might not prioritize or recognize bias concerns. For example, research shows that, in hiring, the gender-specific wording in a job description can reinforce gender inequalities (Gaucher, Friesen, and Kay, 2011). Similar phenomena could occur when selecting vendors. The assessment criteria should be validated for equity across the range of all potential

respondents and solutions that could be delivered, according to the acquisition process's goals. Government requests for proposals could include specific antibias requirements, such as performance and testing requirements in verifying and validating any algorithms.

Second, as in the acquisition planning process, the solicitation language can mischaracterize the use case, causing vendors to develop an application that is not well suited to the purpose for which the product would ultimately be deployed.

Third, solicitations can overlook the issue of bias or not require each proposal to include a bias assessment, monitoring, or mitigation plan. Thus, a vendor could unintentionally develop and deliver a product that is biased simply because it is not directed to manage bias. This may be affected by the extent to which the vendor's workforce is more or less representative of or familiar with the target user population, and this could be a consideration in selection. In addition, a solicitation could be explicit about the importance of monitoring for bias but precede a selection process (which can be driven by cost and schedule over other factors) that does not factor in a vendor's plans to test for and mitigate bias.

Development

Perhaps the most crucial stage in acquisition of an ML tool is the actual development of the tool, in that this stage is the one at which algorithmic bias can occur. Generally, the model development process includes designing, building, training, and testing the model in a simulated environment. The vendor can iterate these steps until the tool reaches the performance needed for deployment in a realistic environment. This step is usually the focus of attention in the current literature on bias in ML.

> The language of the solicitation and criteria for selection must be unbiased. The language needs to accurately characterize the use case, and solicitations must require each proposal to include a bias assessment, monitoring, or mitigation plan.

Supervised ML models—models that produce prediction results by learning from an existing data set with results—are the most likely to be deployed in DHS applications because, among ML model types, they are easiest to develop, most readable, and easiest to deploy.[2] Supervised ML models learn by maximizing or improving an accuracy metric (e.g., maximizing the true

positive and true negative rates of a facial recognition model). However, a myopic focus on a goal of accuracy can lead to models that exhibit unwanted behavioral side effects (Soares and Fallenstein, 2015), such as bias. Such biases are most likely to present as lower model accuracy rates (i.e., higher error rates) on specific subsets of the population.

There are different mechanisms by which an ML training process can produce a biased system. These mechanisms implicate violations largely of the statistical regularity assumption underpinning ML models. (Recall from the introduction that the valid use of ML models for decisionmaking requires statistical regularity in both training and deployment. *Statistical regularity* refers to the condition in which the populations subject to ML decisions are statistically identical during both the training and deployment phases—that is, "don't use a model built for one population on a different population.") A few ways in which the assumption can be broken down include the following:

- **sample-size and base rate differences:** ML models are statistical models. They are therefore subject to statistical laws. Basic statistical laws guiding the efficacy of such models provide that, in general, the accuracy (and hence error rates or confidence levels) of an ML model is proportional to the training sample size. As a result, the model's outputs for samples from minority subpopulations will have lower accuracy than for samples from large or majority subpopulations. When the training sets are large enough, such differences might fade. But, in many cases, these differences are meaningful. A related point made in recent theorems is that, when outcome base rates differ across demographics, it is impossible to simultaneously satisfy certain fairness metrics (Berk et al., 2018).[3]
- **unrepresentative training data and distribution shift:** A model trained on one statistical population will perform differently if deployed on a different population. Such shifts in the population statistics between training and deployment can occur for a variety of reasons:
  - systemic changes in how or where the ML model is deployed
  - training on old or historical data whose collection may have been poisoned by systemic biases
  - training on old data that no longer reflect current realities.

  The overall effect of the shift is a proxy population mismatch. The ML model performance

During development, a model must be trained on data that are appropriately representative of the populations for whom it will be deployed.

will be uncalibrated or misaligned for the new decision population and hence exhibit unequal decisionmaking accuracy or error rates among different groups. During development, the ML training process should be monitored to ensure that models are trained on the appropriate populations.

Delivery

In a subsequent step, DHS receives the developed system and may also integrate it into its own infrastructure. This step can involve verification and validation, including testing the product for performance. At this step, if DHS does *not* include testing for bias in its verification and validation process, it might fail to identify a bias that, when the system is deployed, causes harm. Testing during delivery should therefore include verification and validation to ensure that the product does not contain bias both before and after integration and that processes are in place to detect bias during postintegration use.

Consider, for example, the two applications described under the "Acquisition Planning" section earlier: facial recognition to identify domestic terrorists and a visa-overstay risk tool. A red-teaming of the former might involve testing with a large and diverse test data set of images of people with different demographic characteristics and even different lighting and environmental conditions. A related effort on a risk tool might take historical immigration data reaching several years back to test how the system performs along key demographic and other characteristics. Evidence of bias could be taken back to the vendor and improvements made under the original vendor contract.

Deployment, Maintenance, and Sustainment

Last, the product is deployed for operation, and the client component takes over maintenance and sustainment of the system. Despite the system being deployed, it is far from being finished. Instead, the system is (or should be) monitored to identify and address errors, improve performance, and, as appropriate, build in new features. Because fielding and deployment of a system are often the threshold for measuring success of an acquisition process, significant effort and costs go into maintaining and sustaining it (Gupta, 1983).

There are several mechanisms by which a system's deployment, maintenance, and sustainment can be

Testing during delivery should include verification and validation to ensure that the product does not contain bias—and that it will not do so after integration.

biased. These mechanisms implicate violations largely of the contextual regularity assumption underpinning ML models. (Recall that *contextual regularity* refers to the condition in which the decisionmaking context remains identical between training and deployment—that is, "don't use a model built for one decision for a separate decision, however related.") Examples of these mechanisms include the following:

- **decision model misuse:** A model trained for a specific decision can be uncalibrated for other decisions. This decision context creep occurs occasionally (e.g., the reported use of a recidivism risk estimation tool for informing decisions of sentence length). There is the further question of what decisions it may or may not be prudent to automate with ML models. Various factors enter into this decision: Is the goal merely decision efficiency? Do stakeholders need decision explanations? Are the ML models trusted in that domain? Is the decision high stakes? Is there meaningful human oversight in case of mistakes? Are there relevant, *unbiased*, historical data with which to train the model?
- **automation bias:** A key value of ML decision aids is their ability to off-load expensive cognitive processing for decisionmaking. Over time, that off-loading leads to patterns of decisionmaking that increasingly and uncritically rely on such aids. Unfortunately, without careful design, deployed ML models can "fail silently" (i.e., they give no indications of when their outputs should not be trusted, unlike, for example, failures in physical systems, such as deteriorating infrastructure).
- **unequal application of discretion:** Another source of bias in this step is the unequal application of discretion. Artificial decision aids are often deployed with direct human oversight. This makes for more robustness in decisionmaking in that the human can step in to adjudicate or correct edge cases on which the model might fail. Such human oversight involves an amount of discretion that can be subject to implicit or explicit biases. These can lead to decisionmaking having

Decisions for which a model is trained must be consistent with those it will actually make once deployed. The model must also be monitored to ensure equitability in application and modified as needed throughout its lifetime.

disparate impact on protected and vulnerable groups.
- **model modifications:** After a model is deployed, it can be modified through software updates or patches. Following any such updates, an ML decision aid must be reevaluated to identify potential bias. But the time between software upgrades can vary greatly, and conducting postupdate evaluations can involve significant costs and effort. Continual monitoring for bias could provide early insight into potential issues that can be addressed early on. Including ongoing auditing procedures in deployment documents could help ensure that DHS programs actually mitigate the misuses and biases described.

In sum, bias can be influenced and introduced in several places in an acquisition process, and understanding and mitigating bias requires attention throughout that life cycle.

Opportunity Areas to Address Bias in Machine Learning Software Acquisition

In this exploratory work, we sought to help frame DHS's thinking about the nature of bias in the operational use of ML in LE decisionmaking. The preceding sections introduced a notional ML product acquisition framework to help identify points at which eventual implementations of algorithms can become biased across the course of a typical acquisition process. We identified several points in the acquisition process at which such biases can develop.

The framework needs further tailoring to the operational need and nuances of the various DHS agencies and stakeholders. Stakeholders interested in using algorithmic decision aids should seek to further identify and understand the specific ways in which bias can occur and what domain-specific actions would help mitigate them. For example, when acquiring decision aids for LE purposes, DHS might want to ensure that algorithms treat subgroups in the deployment population in an unbiased fashion. Special care would be needed to neutralize any bias introduced by historical LE practices and to avoid unfairly targeting specific groups (Philimon, 2020). This factor might be important to consider across DHS components.

As various components of DHS acquire technologies with ML capabilities, some considerations will apply broadly, while others will be more specific to one component or another. Mitigation considerations for specific DHS components include the following (some of which they may already be doing):

- **the Transportation Security Administration (TSA):** When acquiring facial recognition or body scanners for LE and surveillance, TSA could adopt various ways to rigorously test such systems to prevent unexpected biases in deployment. Although TSA might not be able to collect demographic data on those going through airports, it might be able to conduct testing that could account for preexisting differences in the composition of populations going through the various airports in the country, which likely vary by

region. Different airports might require different ways to prevent bias.

- **U.S. Customs and Border Protection (CBP):** When acquiring algorithmic decision aids for surveillance, CBP could account for the differences in the composition of populations at the borders and those within the United States. That factor could affect current and future surveillance methods and the introduction of bias into any system.
- **the Federal Emergency Management Agency (FEMA):** When acquiring algorithmic decision aids for disaster response and relief, FEMA should explicitly include selection criteria for equity in the acquisition process (if it does not do so already). For example, FEMA might want to determine whether certain types of disasters or regions tend to receive shorter response times or more aid. In some cases, such disparity might be necessary because of the level of destruction, but, in other cases, it could signal a bias in prioritization of disaster management and reduction.

Opportunity Areas Specific to the Department of Homeland Security

Several DHS-specific crosscutting standards or concerns may warrant attention as DHS increasingly fields operational ML decision aids. We discuss some of them in this section.

Establish Standards and Baselines to Measure Bias in Machine Learning Used by Law Enforcement Agencies

As LE adopts more types of algorithmic decision aids for its uses, a key factor to keep in mind is the standards and baselines used to measure bias. For example, there are differences between models that make decisions about *people* and those that make decisions about *objects*. Both types of models will be vulnerable to containing bias. Models that identify objects, such as those that detect suspicious objects in suitcases at airport security, could have a bias to label toothbrushes as knives. Because most people carry toothbrushes, such a bias would affect travelers regardless of any distinguishing characteristics. This bias would thus incur mainly an economic cost in time spent on unnecessary searches. In contrast, a model biased against a certain type of person can adversely affect people and potentially even their communities. For instance, a model containing a bias that flags residents of certain ZIP Codes as suspicious or high risk would end up targeting specific subsets of the population. This could lead to legal concerns and could reinforce stereotypes against this population group. As a result, the costs of biased decisions on people often go beyond economic into social and psychological costs.

DHS components that, like TSA, use multiple types of models could establish baselines for bias in software. This includes thinking carefully about what dimensions of bias are relevant and how to monitor for bias in their models on those dimensions.[4] One could determine how they currently acquire algorithmic decision aids and how to improve on current standards (or develop new ones) for detecting and mitigating bias. Identification of

when to trigger checks for bias would also be important; the most crucial would likely be comparing models that distinguish among objects and those that distinguish among people.

Identify How to Account for and Weigh All Costs of Biased Outcomes

There are two main kinds of errors for classification ML models: false positives and false negatives. False positives are data points that have incorrectly been labeled positive (e.g., a model detecting for blue cats incorrectly labels a blue dog as a blue cat). False negatives are data points that have incorrectly been labeled negative.

Although the ideal is to reduce both types of error, after a certain point, that is impossible. Often, attempting to lower the false-positive rate ends up increasing the false-negative rate and vice versa. In such cases, it is important to understand what both types of error mean for the model at hand and for the people or objects under scrutiny. Normally, one type of error is more hazardous in a particular context than the other. As such, it is important to be explicit in weighing the costs of both types of error and account for such disproportionate outcomes.

In the example of facial recognition in LE, this balance between false negatives and false positives is crucial given that all errors could result in harm. At first glance, a false negative in this case could be seen as having a higher cost than a false positive: A false positive could lead, for instance, to an unnecessary strip search, whereas a false negative (e.g., failure to detect a weapon) could lead to a more catastrophic result. A false positive might seem to have a lower cost than the result of a false negative labeling. On the other hand, a high rate of false positives, leading to a high number of unfruitful searches, could erode public trust. It is up to LE and other government agencies, which are already working hard to protect the public, to find this balance as they incorporate ML models into their operations.

Provide Workforce Development for Artificial Intelligence Capabilities

As AI systems are increasingly integrated into LE functions, agencies should consider how to train and shape the workforce that will acquire and work with the software, such as reviewing its results. This could include training acquisition personnel who write and evaluate solicitations, or who work with vendors, providing them with ways to avoid introducing bias. Users of ML-based decision aids in the DHS components will also need training on taking into account common limitations of algorithmic decisionmaking, such as considering the impact of implicit value judgments in algorithmic outcomes, understanding how and why ML models can fail, and maintaining backup processes for double-checking possibly incorrect automated decisions. Further, there are no easy fixes for workforce training: Agencies will need to reckon with evidence of mixed effectiveness of workplace diversity (e.g., implicit bias) training (Bezrukova et al., 2016; Kalinoski et al., 2013; Dobbin and Kalev, 2016). In other words, training should be much more targeted than on simply compliance or inclusivity.

Opportunity Areas Not Specific to the Department of Homeland Security

Several courses of action can apply more broadly than any DHS effort to deploy ML-based decision aids.[5] These courses of action reflect generally effective ML deployment practices:

- **certification labels on ML resources:** ML models should have certification labels with information on ML model characteristics (e.g., model purpose and limitation, accuracy and error rates, disaggregated evaluation statistics).[6] Data sets used for training models also need to be clearly characterized for distribution statistics and limitations. Label requirements should be standardized across DHS components. Every ML model deployed by a DHS component should also have a routine scheduled recertification process. The standards for such certification should be guided by neutral or independent parties.
- **performance tracking and disaggregated evaluation:** Clear, consistent processes are required to continuously measure the performance of deployed ML decision aids. This includes evaluating and tracking the accuracy of the ML model (e.g., false positives and false negatives). The process also includes repeated audits for bias. Bias audits include disaggregated evaluation of ML models (e.g., measure model performance on different, relevant subdemographics; see Mitchell et al., 2019). The need for tracking can place specific demands on operations. For example, one might need to implement experimental design setups to try to estimate traditionally unobserved signals (e.g., false negatives, sensitive attributes of individuals).
- **impact assessments and continuous red-teaming:** ML models should undergo extensive checks at each step to ensure quality control. This could be in the form of *red-teaming* or *security tests*. Where feasible, the use of model implementations can also enable more-robust external verification. The continuous application of red-teaming enables ongoing estimation of operational robustness to new ML security threats. DHS should also require an algorithmic impact assessment (Reisman et al., 2018) when

DHS can avoid harmful bias in its machine learning systems by establishing standards for measuring bias, weighing costs of biased outcomes, and providing workforce development for the emerging technologies.

planning to apply ML-based decision aids. This would include clear articulation of the decision to be automated, minimal performance specifications, and audit methodology. Algorithmic impact assessments should also be produced for any third-party models before integration or deployment.

Ways Forward

Significant further work remains to flesh out the framework presented in this Perspective and to highlight bias entry points in a more DHS-specific context to make the framework more productive for DHS acquisition planning. There are also unanswered research questions that could help DHS leaders think productively about the question of bias in operational ML tools. Each of these could form the basis of further research. These questions include the following:

- **a DHS-specific examination of bias in decision-making:** Our general definition of *bias* did not account for legal and other obligations, such as antidiscrimination. What would be a practical definition of *ML equity* or *bias* for DHS? How would that definition vary across the different kinds of DHS services and operations? What measure and audit procedures would address these bias definitions?
- **developing best practices on ML decision automation and bias audits:** What kinds of operations within the components are most likely to benefit from automation with ML tools? What kind of automation would be relevant (e.g., fully automated, human-in-the-loop)? And which of these automatable operations are especially susceptible to bias concerns? What are specific bias mitigation strategies? Alternatively, what kinds of decisions and operations *should not* be automated? And for what reasons?
- **developing best practices on reducing bias in ML tool acquisitions:** What are effective ways to specify contracts for third-party ML tool development to reduce bias? What antibias standards should DHS adopt for verifying, validating, and integrating ML acquisitions? What is the life-cycle management framework for these tools? What should the relevant model recertification processes look like?

Appendix. Bias Concerns in the Notional Acquisition Framework

The table describes examples of potential ways in which bias can be introduced at various points in the software acquisition life cycle. For each stage (e.g., acquisition planning), we indicate several examples of different types of bias (e.g., bias in underlying social context that is underappreciated) that could arise , and suggested mitigation steps, within that stage.

Examples of Bias and Mitigation Steps at Various Points in the Acquisition Process

	Acquisition planning	Solicitation and selection	Development	Delivery	Deployment, maintenance, and sustainment
Bias	• Bias in the underlying social context could be under-appreciated. • The use case could be mischaracterized or under-characterized. • Value judgments about different outcomes from an AI or ML system could remain implicit. • Acquisition teams might not share or understand the characteristics of the people the AI or ML system would affect.	• The use case could be mischaracterized or undercharacterized. • Solicitations could overlook bias. • Vendors' plans for avoiding bias might not be factored into the selection process.	• Assumptions behind the model could be biased. • Training or test data set samples could be unbalanced. • The costs of different kinds of errors could be incorrectly estimated. • The training data could be unrepresentative of the affected population or contain historical bias.	• Verification and validation might not include testing for bias. • Models might not be labeled with specific use-case or decision purpose and training data statistics. • The product could lack any or adequate explanation of contexts under which the model can go wrong.	• There could be bias in where the system is deployed. • There could be a distribution shift in the real-world or deployed population sample. • There might be no monitoring of performance or shifts in deployed accuracy.
Mitigation steps	• (Acquisition team) Understand the social context in which the technology will be used. • Deliberate and identify the human impact of biased outcomes, and make explicit any subjective judgments of outcomes. • Ensure that the acquisition team is representative of the populations for whom the technology is intended. • Include bias-mitigation requirements and provisions in capability needs, concepts of operations, requirements, and design documents, such as in the Joint Requirements Integration and Management System process.	• In requests for proposals include specific antibias requirements, such as performance and testing requirements in verifying and validating any algorithms.	• Monitor the ML training to ensure that models are trained on the appropriate populations.	• Conduct red-teaming, such as testing with a large and diverse test data or historical data. • Collect evidence of bias so that the vendor can make improvements under the original vendor contract.	• Continually monitor for bias to provide early insight into potential issues that can be addressed early on. • Include ongoing auditing procedures in deployment documents to help ensure that DHS programs actually mitigate the misuses and biases described.

Notes

[1] COMPAS was developed by a company called Northpointe, which later became Equivant.

[2] The same general points hold for unsupervised and reinforcement learning models (the other main types of ML models). But there will be special nuances for each of these other types.

[3] This can happen when there is a correlation between group membership and the outcome or target of prediction. For example, if people from a given population group are more likely to have a certain outcome, a good ML model would need to discriminate on the basis of that demographic attribute. That would fail to maintain some equity metrics, such as demographic parity.

[4] Such dimensions might be, for instance, those derived from ethical or legal mandates for equity.

[5] These assume an operating model in which DHS and its components use ML models developed by third parties subject to contracted terms, as opposed to one in which ML models are internally developed.

[6] *Disaggregated evaluation* refers to the practice of evaluating the performance of a decisionmaking model using samples from disaggregated subpopulations. This helps users measure or observe any variation in performance that is correlated to demographic or other sensitive attributes, such as whether the model is less accurate for black women than for the background population.

Bibliography

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias," *ProPublica*, May 23, 2016. As of May 1, 2019: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth, "Fairness in Criminal Justice Risk Assessments: The State of the Art," *Sociological Methods and Research*, July 2018.

Bezrukova, Katerina, Chester S. Spell, Jamie L. Perry, and Karen A. Jehn, "A Meta-Analytical Integration of over 40 Years of Research on Diversity Training Evaluation," *Psychological Bulletin*, Vol. 142, No. 11, 2016, pp. 1227–1274.

Brenan, Megan, "Amid Pandemic, Confidence in Key U.S. Institutions Surges," Gallup, August 12, 2020. As of September 29, 2020: https://news.gallup.com/poll/317135/amid-pandemic-confidence-key-institutions-surges.aspx

Chouldechova, Alexandra, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments," *Big Data*, Vol. 5, No. 2, June 1, 2017, pp. 153–163.

Code of Federal Regulations, Title 48, Federal Acquisition Regulations System; Chapter 1, Federal Acquisition Regulation; Subchapter A, General; Part 2, Definitions of Words and Terms; Subpart 2.1, Definitions; Section 2.101, Definitions. As of September 29, 2020: https://www.ecfr.gov/cgi-bin/text-idx?SID=0e40a40b55b199ede3cca0052fc88d30&mc=true&node=se48.1.2_1101&rgn=div8

DHS—*See* U.S. Department of Homeland Security.

Dieterich, William, Christina Mendoza, and Tim Brennan, *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*, Northpointe Research Department, July 8, 2016. As of May 8, 2019: http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf

Dobbin, Frank, and Alexandra Kalev, "Why Diversity Programs Fail," *Harvard Business Review*, Vol. 94, No. 7–8, July–August 2016, pp. 52–60.

Donini, Michele, Luca Oneto, Shai Ben-David, John S. Shawe-Taylor, and Massimiliano Pontil, "Empirical Risk Minimization Under Fairness Constraints," in S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds., *Advances in Neural Information Processing Systems*, No. 31, 2018, pp. 2791–2801. As of December 7, 2020: https://papers.nips.cc/paper/2018/hash/83cdcec08fbf90370fcf53bdd56604ff-Abstract.html

Ferguson, Andrew G., "Policing Predictive Policing," *Washington University Law Review*, Vol. 94, No. 5, 2017, pp. 1109–1189. As of December 7, 2020: https://openscholarship.wustl.edu/law_lawreview/vol94/iss5/5/

Gaucher, Danielle, Justin Friesen, and Aaron C. Kay, "Evidence That Gendered Wording in Job Advertisements Exists and Sustains Gender Inequality," *Journal of Personality and Social Psychology*, Vol. 101, No. 1, January 2011, pp. 109–128.

Groves, Kevin S., and Ann E. Feyerherm, "Leader Cultural Intelligence in Context: Testing the Moderating Effects of Team Cultural Diversity on Leader and Team Performance," *Group and Organization Management*, Vol. 36, No. 5, 2011, pp. 535–566.

Gupta, Yash P., "Life Cycle Cost Models and Associated Uncertainties," in J. K. Skwirzynski, ed., *Electronic Systems Effectiveness and Life Cycle Costing*, Berlin: Springer, North Atlantic Treaty Organization Advanced Science Institutes Series F, Computer and Systems Sciences, Vol. 3, 1983, pp. 535–549.

Hollywood, John S., Andrew Lauland, Dulani Woods, Kenneth N. McKay, and Yingzi Zhang, *Better Policing Toolkit*, Santa Monica, Calif.: RAND Corporation, TL-261-RC, 2018. As of September 29, 2020:
https://www.rand.org/pubs/tools/TL261.html

Kalinoski, Zachary T., Debra Steele-Johnson, Elizabeth J. Peyton, Keith A. Leas, Julie Steinke, and Nathan A. Bowling, "A Meta-Analytic Evaluation of Diversity Training Outcomes," *Journal of Organizational Behavior*, Vol. 34, No. 8, 2013, pp. 1076–1104.

Liptak, Adam, "Civil Rights Law Protects Gay and Transgender Workers, Supreme Court Rules," *New York Times*, June 15, 2020. As of June 16, 2020:
https://www.nytimes.com/2020/06/15/us/gay-transgender-workers-supreme-court.html

Mazerolle, Lorraine, Sarah Bennett, Jacqueline Davis, Elise Sargeant, and Matthew Manning, "Procedural Justice and Police Legitimacy: A Systematic Review of the Research Evidence," *Journal of Experimental Criminology*, Vol. 9, No. 3, 2013, pp. 245–274.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru, "Model Cards for Model Reporting," in *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency*, New York: Association for Computing Machinery, 2019, pp. 220–229.

National Research Council, *The Growth of Incarceration in the United States: Exploring Causes and Consequences*, Washington, D.C.: National Academies Press, 2014.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan, "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science*, Vol. 366, No. 6464, October 25, 2019, pp. 447–453.

Office of the Under Secretary for Acquisition and Sustainment, U.S. Department of Defense, "Operation of the Adaptive Acquisition Framework," Department of Defense Instruction 5000.02, January 23, 2020. As of September 29, 2020:
https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/500002p.pdf?ver=2020-01-23-144114-093

Office of the Under Secretary for Management, U.S. Department of Homeland Security, "Acquisition Management Directive," Directive 102-01, rev. 03.1, July 28, 2015, incorporating change 1, February 25, 2019. As of October 27, 2020:
https://www.dhs.gov/sites/default/files/publications/mgmt/planning-and-budgeting/mgmt-dir_102-01-acquisition-management-directive_revision-03-1.pdf

Oliver, David, "Facial Recognition Scanners Are Already at Some US Airports: Here's What to Know," *USA Today*, August 16, 2019. As of September 29, 2020:
https://www.usatoday.com/story/travel/airline-news/2019/08/16/biometric-airport-screening-facial-recognition-everything-you-need-know/1998749001/

Osoba, Osonde A., Benjamin Boudreaux, Jessica Saunders, J. Luke Irwin, Pam A. Mueller, and Samantha Cherney, *Algorithmic Equity: A Framework for Social Applications*, Santa Monica, Calif.: RAND Corporation, RR-2708-RC, 2019. As of September 29, 2020:
https://www.rand.org/pubs/research_reports/RR2708.html

Osoba, Osonde A., Benjamin Boudreaux, and Douglas Yeung, "Steps Towards Value-Aligned Systems," in *AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, New York: Association for Computing Machinery, 2020, pp. 332–336.

Philimon, Wenei, "Not Just George Floyd: Police Departments Have 400-Year History of Racism," *USA Today*, June 7, 2020. As of November 6, 2020:
https://www.usatoday.com/story/news/nation/2020/06/07/black-lives-matters-police-departments-have-long-history-racism/3128167001/

Public Law 88-352, Civil Rights Act of 1964, July 2, 1964. As of September 29, 2020: https://www.govinfo.gov/content/pkg/STATUTE-78/pdf/STATUTE-78-Pg241.pdf

Raji, Inioluwa Deborah, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes, "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, January 2020, pp. 33–44.

Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker, "Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability," *AI Now Institute*, 2018, pp. 1–22. As of September 29, 2020: https://ainowinstitute.org/aiareport2018.pdf

Soares, Nate, and Benja Fallenstein, "Aligning Superintelligence with Human Interests: A Technical Research Agenda," Berkeley, Calif.: Machine Intelligence Research Institute, Technical Report 8, 2015.

Todd, Andrew R., Galen V. Bodenhausen, Jennifer A. Richeson, and Adam D. Galinsky, "Perspective Taking Combats Automatic Expressions of Racial Bias," *Journal of Personality and Social Psychology*, Vol. 100, No. 6, 2011, pp. 1027–1042.

U.S. Code, Title 5, Government Organization and Employees; Part I, The Agencies Generally; Chapter 5, Administrative Procedure; Subchapter II, Administrative Procedure; Section 552a, Records Maintained on Individuals. As of December 7, 2020: https://www.govinfo.gov/app/details/USCODE-2010-title5/USCODE-2010-title5-partI-chap5-subchapII-sec552a

U.S. Department of Homeland Security, "Snapshot: Public Safety Agencies Pilot Artificial Intelligence to Aid in First Response," press release, October 16, 2018. As of September 29, 2020: https://www.dhs.gov/science-and-technology/news/2018/10/16/snapshot-public-safety-agencies-pilot-artificial-intelligence

U.S. National Archives and Records Administration, "EEO Terminology," last reviewed August 15, 2016. As of November 6, 2020: https://www.archives.gov/eeo/terminology.html

Williams, Timothy, "Study Supports Suspicion That Police Are More Likely to Use Force on Blacks," *New York Times*, July 7, 2016.

Acknowledgments

We are grateful to Emma Westerman and Terrence Kelly from the Homeland Security Operational Analysis Center for their guidance and support of this research. We also thank the subject-matter experts from the RAND Corporation and the U.S. Department of Homeland Security who provided helpful insight. Finally, we thank our reviewers, Victoria Greenfield, John Hollywood, William Shelton, and William Welser IV, for helping to improve this Perspective.

About the Authors

Douglas Yeung is a social psychologist at the RAND Corporation and a member of the Pardee RAND Graduate School faculty. In his research, he has explored the use of emerging technologies (e.g., social media, mobile devices) for policymaking. He has a PhD in psychology.

Inez Khan is a research assistant at the RAND Corporation. She is interested in interdisciplinary research on applying machine learning (ML) methods to the public policy space. She has a BS in statistics and ML with minors in physics and international relations.

Nidhi Kalra is a senior information scientist at the RAND Corporation. Her research focuses on autonomous vehicle policy, climate change adaptation, and tools and methods that help people and organizations make better decisions amid deep uncertainty. She has a PhD in robotics.

Osonde A. Osoba is a senior information scientist at the RAND Corporation and a professor at the Pardee RAND Graduate School. His recent applied work includes ML modeling for strategic decision support, reinforcement learning applied to policy-relevant agent-based models, and causal modeling of social and behavioral phenomena; his recent work on the implications of artificial intelligence and ML has focused on issues of data privacy and fairness in artificial intelligence and algorithmic decision systems generally. He has a PhD in electrical engineering.

About This Perspective

As law enforcement (LE) agencies increasingly integrate machine learning (ML) into diverse operations, there is value in understanding the shortcomings of decision aids based on ML technologies. This Perspective focuses on a specific shortcoming of such ML-based decision aids: the multiple types of bias that can result in disparate harms in LE applications. It is important to understand any challenges that exist in the fairness, efficiency, and effectiveness of these individual systems. This Perspective is intended to identify opportunity areas that can help improve U.S. Department of Homeland Security acquisition of ML-based decision aids. It might also be of interest to LE agencies that are either fielding or planning to field such tools. Specifically, this work describes the potential for acquisition and implementation biases that go beyond biases that occur in the development of ML models.

This research was conducted using internal funding generated from operations of the RAND Homeland Security Research Division (HSRD) and within the HSRD Acquisition and Development Program. HSRD conducts research and analysis for the U.S. homeland security enterprise and serves as the platform by which RAND communicates relevant research from across its units with the broader homeland security enterprise.

For more information on the RAND Acquisition and Development Program, see http://www.rand.org/hsrd/hsoac or contact the program director (contact information is provided on the webpage).

www.rand.org

PE-A862-1